ORIGINAL PAPER

Adding value to cocoa (*Theobroma cacao* L.) germplasm information with domestication history and admixture mapping

**Maria Marcano · Tatiana Pugh · Emile Cros · Sonia Morales · Elvis A. Portillo Páez ·
Brigitte Courtois · Jean Christophe Glaszmann · Jan M. M. Engels · Wilbert Phillips ·
Carlos Astorga · Ange Marie Risterucci · Olivier Fouet · Ventura González ·
Kai Rosenberg · Isabelle Vallat · Manuel Dagert · Claire Lanaud**

Abstract A sound understanding of crop history can provide the basis for deriving novel genetic information through admixture mapping. We confirmed this, by using characterization data from an international collection of cocoa, collected 25 years ago, and from a contemporary plantation. We focus on the trees derived from three centuries of admixture between Meso-American Criollo and South American Forastero genomes. In both cacao sets of individuals, linkage disequilibrium extended over long genetic distances along chromosome regions, as expected in populations derived from recent admixture. Based on loose genome scans, genomic regions involved in useful traits were identified. Fifteen genomic regions involved in seed and fruit weight variation were highlighted. They correspond to ten previously identified QTLs and five novel ones. Admixture mapping can help to add value to genetic resources and thus, help to encourage investment in their conservation.

M. Marcano (✉) · T. Pugh · E. Cros · E. A. Portillo Páez ·
B. Courtois · J. C. Glaszmann · A. M. Risterucci ·
O. Fouet · C. Lanaud
Centre de Coopération Internationale en Recherche
Agronomique pour le Développement, (CIRAD) TA 40/03,
34398 Montpellier Cedex 5, France
e-mail: marseg@ula.ve

M. Marcano · S. Morales · M. Dagert
Laboratorio de Genética y Química Celular
(GEQUIMCEL), Universidad de los Andes (ULA),
La Hechicera, Mérida, Edo. Mérida, Venezuela

T. Pugh
Facultad de Agronomía,
Universidad Central de Venezuela (UCV),
Av. El Limón, Maracay, 2101 Edo. Aragua,
Venezuela

E. A. Portillo Páez
Facultad de Agronomía. Maracaibo,
Universidad del Zulia (LUZ), Edo Zulia, Venezuela

Introduction

Crop genetic resources provide the foundation for sustaining agricultural production. For most crops, large germplasm collections have been assembled and partially characterised for useful traits (FAO 1998). However, crop-breeding programs can incorporate only a small part

J. M. M. Engels
Bioversity International,
Via dei Tre Denari, 472a, 00057,
Maccarese, Rome, Italy

W. Phillips · C. Astorga
Centro Agronómico Tropical de Investigación y Enseñanza,
P.O. Box 7170, CATIE, Turrialba, Costa Rica

V. González
Instituto Nacional de Investigaciones Agrícolas,
INIA–CENIAP, Av. Universidad, Maracay,
2101 Edo. Aragua, Venezuela

K. Rosenberg · I. Vallat
VALRHONA, Les lots, 26600 Tain L' hermitage, France

of the available diversity at any given time. The long-term maintenance of such collections is a financial burden for those responsible, particularly for long-lived tree species whose seeds are unsuited to storage. Drawing more immediate value from such collections is essential to ensure institutional commitment to their management.

Association mapping has recently emerged as a powerful method for generating genetic information from population samples. It is based on the non-random association of alleles, at two or more loci, called "linkage disequilibrium" (LD), in natural or cultivated populations. Through LD, the genes involved in the variation of morphological, agronomic and other phenotypic traits can be located in the genome if statistical associations can be found between the traits and genetic marker polymorphisms.

Although association mapping has been extensively used for mapping disease factors in humans, its use in plants has only recently begun. Associations were found between polymorphisms at the DNA sequence level between the maize *Dwarf8* gene implicated in plant height, and flowering time, in 92 inbred lines (Thornsberry et al. 2001). In *Arabidopsis*, associations between genotype and phenotype were reported for four known genes involved in flowering time and pathogen resistance in a study that included sequenced fragments genomewide distributed every 100 Kb (Aranzana et al. 2005). Recently, associations were found in a collection of modern barley cultivars, between AFLP markers and mean yield and yield stability (Kraakman 2004). In cacao, an association study was reported in a population which included productive and non-productive clones in which 13 associated markers were near previously mapped QTLs for productivity traits (Schnell et al. 2005).

During their evolution, allopatric populations are subjected to parallel drift and contrasted selection pressures, and can accumulate mutations, generating global divergence. Such structuring of the overall population into independent subpopulations can generate a pattern of LD that typically associates regions distributed in the whole genome. Domestication selected only specific forms of the wild ancestors, and modern breeding focused only on a few parental genotypes. Both processes constituted considerable genetic bottlenecks in crop populations, further accentuating LD. When migrating humans carried crop cultivars with them, recombination events between previously divergent populations sometimes occurred. Thus, LD was reduced between unlinked genome regions but was maintained within segments little affected by recombination; "Global" LD was dissipated or disappeared whereas "local" LD was maintained. LD enables appli-cation of association mapping when no candidate gene is available for the traits under consideration.

When genotype information is available of the founding population, such recent admixture of previously diverged populations can be exploited for mapping purposes in a way that is similar to cross based linkage mapping. Markers can be used to trace chromosomal blocks back to one of the founding populations and QTLs that have diverged between these founding populations can be identified via trait ancestry associations in the admixed populations (McKeigue 2005).

Admixture mapping has been applied to detect genetic factors that determine hypertension, a complex human disease (Zhu et al. 2005), however, to our knowledge, it has not been exploited in plant populations. The approach requires generally a lower density of markers than for traditional association mapping and is particularly useful in admixed populations derived from two known progenitors, with different proportions of the allele affecting the trait (Darvasi and Shifman 2005)

We explored the value of this approach for analysing cocoa germplasm populations, taking advantage of a particular hybrid group that has been recently domesticated.

Theobroma cacao L. is a perennial tree native to the American tropics; it was domesticated more than 2,000 years ago by Mesoamerican peoples, who cultivated a high-quality aromatic chocolate variety, named Criollo (local), probably originating from the northern part of South America (Motamayor et al. 2002). After Spanish colonization, the production of Criollo spread to South America and the Caribbean region to satisfy an increasing demand for cocoa in Europe. A natural disaster in Trinidad during the 18th century led to the introduction to this island of a newly domesticated and genetically divergent variety, a Forastero (foreigner) originating from Lower Amazonia. Open pollinations involving the remaining Criollo trees and the newly introduced Forastero, gave rise to hybrid forms called Trinitario. Due to their vigour, Trinitario materials were later introduced to the South American continent, and gradually spread into original Criollo plantations, leading to further crossing between Criollo and Trinitario individuals. Most of the modern Criollo varieties, selected for their quality traits, result from the recombination between ancestral Criollo and invading Trinitario (Motamayor et al. 2003).

We have recently demonstrated that 80% of the Trinitario trees originated from a very narrow genetic base represented by an almost monomorphic ancestral Criollo form, and a small number of inter-related Forastero types (Motamayor et al. 2003). The first recombination between the two ancestral forms

occurred 250 years ago. Since then, probably no more than 6 or 7 recombination generations have occurred. Modern Criollo/Trinitario varieties now constitute the basis of nearly 70% of cocoa cultivation worldwide.

Although quality and morphology of beans are the most distinctive traits between ancient Criollo and other cacao varieties, additional traits, such as growth habit, pigmentation in different plant structures, yield potential and disease resistance, differentiate ancient Criollo from Forastero cultivars.

In this study, our objectives were to analyze and compare LD in two sets of cacao individuals: a collection and a contemporary plantation, to evaluate the structure in both populations and the feasibility of association genetics to identify regions implicated in two domestication traits.

Materials and methods

Phenotype evaluation

We analyzed two groups of individuals. One group (collection) included 150 Criollo/Trinitario varieties from the Criollo/Trinitario section of the international collection maintained by CATIE (Costa Rica). This collection was extensively characterized 25 years ago for a large number of agronomic traits. The bean and fruit weight data used in this study were provided by Engels ([1981](#)) and stored in the international cocoa germplasm database, ICGD (Wadsworth et al. [2003](#)).

The other group (plantation) was derived from a cacao farm near Mérida, Venezuela, that included a mix of typical Criollo/Trinitario forms. A thousand trees were selected on the basis of morphology using flower, seed, pod and tree architecture traits, and were screened with ten microsatellites. This enabled us to choose 291 trees exhibiting the typical alleles of both Criollo and Forastero ancestral forms for further genotyping and morphological characterisation. Pod weight was evaluated as the average weight from 30 pods/tree and seed weight, was the average weight of one seed, from 400 g of fresh beans/tree. Seed and fruit weight, exemplify traits subjected to domestication processes and were observed in both sets of trees, as well as in other classical map-based QTL analyses.

Genotyping

Different sets of markers were used:

1. Twenty independent microsatellite markers distributed on the ten cocoa chromosomes were analyzed in both sets of individuals (collection and plantation) to evaluate population structure and determine the threshold to be used in LD analysis.

2. One hundred and one microsatellites were used to genotype the 150 individuals from the CATIE collection and 92 microsatellites were analyzed in the 291 individuals from the Mérida plantation. These microsatellites were used to study the extent of local LD across the whole genome and to identify marker-trait associations. Microsatellite map positions were based on the last cacao linkage map published (Pugh et al. [2004](#)). DNA extraction was performed from fresh leaves according to Risterucci et al. ([2000](#)). Genotyping was carried out as described by Pugh et al. ([2004](#)).

Thanks to the strong molecular differentiation between the two ancestral forms, marker alleles could be classified as Criollo or Forastero alleles for most loci. This categorization was made with reference to the Criollo ancestral genotype (LAN 1), collected in the Lacandona forest (Mexico) near Mayan ruins (Motamayor et al. [2002](#)), and to the Matina 16 Forastero clone identified as one of the main Forastero ancestors (Motamayor et al. [2003](#)), and was refined using a factorial analysis of correspondences (FAC) (Benzecri [1973](#)) and allele projections on the two first FAC eigenvectors, which discriminated the Criollo and Forastero types (data not shown).

To focus specifically on the presence of the ancestral Criollo chromosomal fragments in these sets of accessions, we considered only two classes of alleles for association studies: Criollo versus non-Criollo alleles. Thus, microsatellite markers were scored as bi-allelic markers.

Data analysis

Population structure analysis

The patterns of population structure in each set of individuals were investigated in two steps: initially with a factorial analysis of correspondence, using GENETIX 4.02 software (Belkhir et al. [2001](#)) and then with the bayesian clustering method implemented in the STRUCTURE program (Pritchard et al. [2000](#)). The latter program assigns individual genotypes to a user-defined number of clusters (K), achieving linkage equilibrium within cluster. The model assumes Hardy–Weinberg equilibrium (HWE) within each subpopulation. We used a model with admixture and uncorrelated allele frequencies. STRUCTURE estimated the proportion of ancestry in each of the K cluster for all individuals.

For each run, a burning period of 250,000 iterations was used and estimates were obtained using 500,000 Markov Chain Monte Carlo repetitions.

Linkage disequilibrium analysis

The composite measure of LD (Δ_{ij}) was used, which is suitable for the phase-unknown situations (Weir and Cockerham 1979). This composite estimator measures the association of alleles from different loci on the same haplotype (intragametic LD) as well as on different haplotypes (intergametic LD). Δ_{ij} was estimated using the Linkdos program (Garnier-Gere and Dillmann 1992) adapted from Black and Krafsur's program (1985) used by GENETIX 4.02 software. To illustrate the extent of LD, we plotted LD values against genetic distance between each pair of microsatellite loci per chromosomal region. Genetic distances were determined with a high-density genetic map previously established (Pugh et al. 2004). The threshold evidencing the existence of LD along the chromosomes was determined in each population with the LD values for 300 pairs of independent markers located on different chromosomes. A LD value which remained below 99% of the observed values was chosen as threshold in each population (bold lines in the Fig. 2).

Hardy–Weinberg equilibrium and haplotypes estimation

At every marker loci, genotype proportions were tested for Hardy–Weinberg equilibrium using a permutation version of the exact test given by Guo and Thompson (1992) implemented in Powermarker program (http://www.statgen.ncsu.edu/powermarker). Bootstrapping was performed (100 bootstraps). Sequential Bonferroni adjustments were used to determine statistical significance (Rice 1989). In addition, we also used this program to calculate the haplotypes frequencies using the expectation-maximization (EM) algorithm, an iterative method to reconstruct haplotypes and find frequencies to maximize the likelihood of the genotype data (Excoffier and Slatkin 1995). The EM algorithm is based on the assumption that genotype frequencies at each locus are in HWE. We estimated the haplotypes in those regions where closely linked loci showing significant association were in Hardy–Weinberg equilibrium.

Molecular marker-trait associations

For each set of cacao plants, Normality of traits was tested with the Shapiro–Wilk test and variance homogeneity with the Levene test by the UNIVARIATE and GLM Procedure of SAS, respectively (SAS Institute Inc., 1998). Marker data were compared with trait values by three methods: a one-way analysis of variance (ANOVA) performed with SAS, a non-parametric Kruskal–Wallis test and a likelihood of the odds (LOD) score test both performed by MapQTL 4 (Van Ooijen and Maliepaard 1996). For the ANOVA test, the marker-trait association was declared significant for $P \leq 0.005$. To establish the LOD critical values, a permutation test with 1,000 replications was conducted using MapQTL 4. The LOD threshold value for each individual test corresponding to 5% significance at the genome level was fixed at 2.5. LOD scores above this level were considered to be significant.

Results

Structure and diversity of both sets of accessions

In a first round of analyses, the population structure of the collection and the plantation samples were checked individually with a FAC. In all cases where several individuals appeared clustered on the first plane of the FAC, one of them was arbitrarily selected and the others discarded from the analysis. These clusters grouped genotypes differing only for 0–5% of the loci. They represented 32 and 30 individuals, respectively and were discarded from the analyses.

According to the results obtained with the STRUCTURE software, with 20 independent microsatellite markers, the most probable number of distinct sub populations for interpreting the observed genotypes, the K was 1 for the collection and 2 for the plantation. For the Mérida plantation, the Criollo and Forastero genotypes used as standards in the analyses appeared to be the main founders of these two sub-groups, in which each one contributed with the highest possible level, respectively 96 and 97%. A new set of 197 individuals was selected from the 261 individuals of the Mérida plantation after eliminating those individuals with a contribution of membership to one of the two sub-groups above 90%. An additional analysis with STRUCTURE showed continuous variability in the new sub-set, without distinct sub groups ($K = 1$).

The diversity of the two datasets (118 individuals for the collection and 197 for the plantation) is summarized in Fig. 1, which results from a FAC including 64 loci common to both groups of individuals. The datasets are complementary in terms of the variability relating to the two main ancestors: The sample from the CATIE collection is closer to the Forastero founder
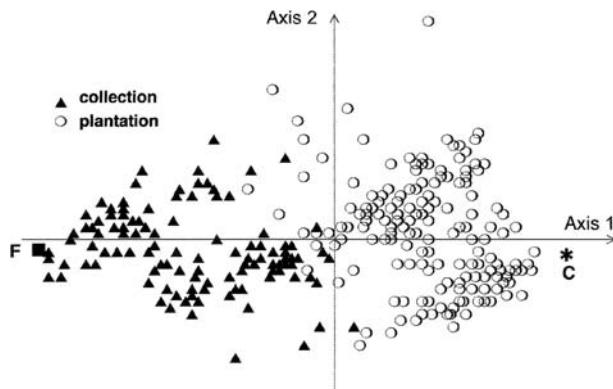
Fig. 1 Diversity of 315 Criollo/Trinitario accessions from a collection and a plantation and their potential ancestors, F■ (Forastero) and C* (Criollo) evaluated with 64 microsatellites. The first plane (axe 1/2) of factorial analysis of correspondence (FAC) explains 39% of the total variation. F and C are located at the extremes of the populations, F being closer to the collection and C to the plantation. Both populations represent complementary hybrid forms between their potential ancestors

whereas the sample from the Mérida plantation is closer to the ancestral Criollo type. The observed levels of heterozygosity are similar 0.51 and 0.44, respectively.

Extent of linkage disequilibrium in both sets of accessions and haplotype diversity

The extent of LD in each set of accessions is shown in Fig. 2, for the chromosomes with the highest number of mapped genetic markers (CH1 and CH9), as plots of LD values (Δ) against genetic distance between each pair of markers (cM). Along each of the chromosomes, LD values decreased as the distance between loci increased, with a similar rate of decay among chromosomes. As a reference for comparison, the distribution of the LD values between 300 pairs of independent

markers (located on different chromosomes) remained below 0.078 and 0.062 for 99% of the values observed in the collection and the plantation, respectively. Taking these values as thresholds, we concluded that LD typically disappeared by 25–35 cM in the sampled individuals.

A trend towards a more rapid decay of LD values was observed for the CATIE collection compared to the Mérida plantation.

In outcrossing species, without access to parents and grandparents, the linkage phase of double heterozygotes cannot be determined directly. Statistical haplotype estimation and evaluation of LD are possible from regions where most markers are in Hardy–Weinberg equilibrium (Weir 1996). In the collection, the most frequent haplotypes were constituted by full-length Forastero, followed by full-length Criollo, allele haplotypes. Similar analyses made on the plantation sample revealed a reciprocal situation where the most frequent haplotypes were constituted by whole Criollo haplotypes followed by whole Forastero haplotypes.

Admixture mapping

Association of markers to seed and fruit weight were consistently detected through the different methods (ANOVA, Kruskall and Wallis non parametric test and maximum likelihood analysis), corresponding to 15 genomic regions, 5 of which were common to both sets of samples (Table 1).

Ten regions (67%) were coincident to QTLs previously identified by classical mapping analysis carried out for the same traits, in four different progenies involving Trinitario (DR1, S52, UF676) or Forastero parents (IMC78, Pound12, Catongo) (Crouzillat et al. 2000; Clément et al. 2003a, b). Five new regions were also associated to the evaluated traits.

Fig. 2 Linkage disequilibrium as a function of genetic distance for markers located in chromosomes 1 and 9. *Bold lines*: threshold for LD for the collection (0.08) and the plantation (0.06). LD typically disappears by 25–35 cM in our cocoa accessions
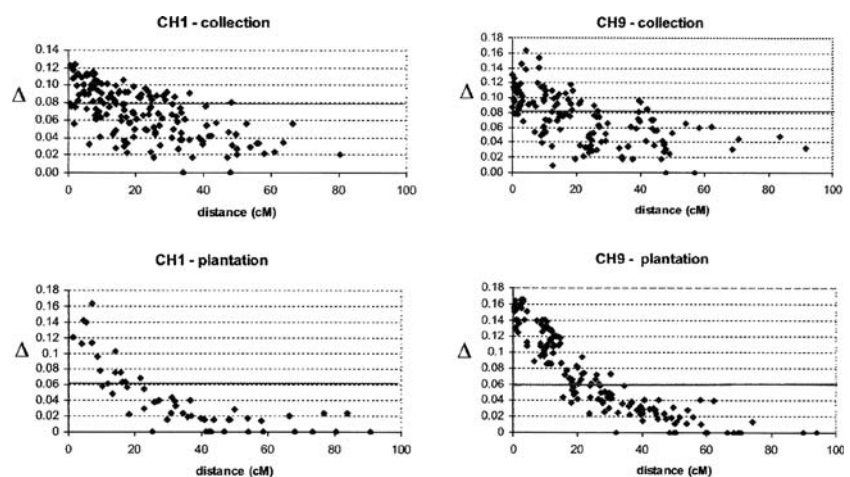
Table 1 Associations between markers and seed and fruit weight variations in Criollo/Trinitario varieties

Trait	Linkage group	Position (cM[a])	Associations LOD score (% explained variability)		QTLs identified by classical methods[b]
			Collection	Plantation	Progeny
Seed weight	1	12–27	3.65 (17.4)	3.94 (11.3)	IMC78
	2	43–58	4.88 (25.0)	6.77 (18.2)	S52-IMC78
	3	51			IMC78
	4	8–24		2.57 (6.5)	DR1
	4	60–75			S52-IMC78-P12
	5	0	3.41 (17.8)		
	5	58–60		2.57 (6.3)	P12
	6	2			UF676
	6	34	3.61 (17.2)		IMC78
	7	1		3.53 (12.7)	
	9	23–36	4.73 (21.9)	2.79 (15.0)	P12
	9	50–60	5.71 (25.9)	2.75 (9.9)	S52-UF676
Fruit weight	1	18			S52-DR1-UF676
	1	95		3.03 (7.6)	
	2	40–52			S52-P12
	3	59			S52
	4	6–19	5.5 (25.4)	2.53 (6.2)	DR1-IMC78
	4	39–59		2.96 (7.6)	P12
	5	15–26			S52-UF676
	6	63		4.25 (10.5)	
	7	50			S52-IMC78
	9	40–50		2.66 (7.6)	P12
	10	72		4.29 (10.8)	

[a] QTL position with 1 LOD support intervals around the peak.

[b] References: Crouzillat et al. (2000), Clément et al. (2003a, b) and Lanaud et al. (2003).

Discussion

We have shown that in cocoa, LD can span long genetic distances along chromosome regions, as expected in populations derived from recent admixture. The results of LD analysis conducted with microsatellite markers indicate that a loose genome coverage, with markers spread every 10–15 cM, is sufficient to identify chromosome regions involved in trait variations in these Criollo/Trinitario accessions.

The tendency to a more rapid decay of the LD values observed for the CATIE collection compared to the Mérida plantation is likely to be the result of a more diverse origin of the germplasm collection, where the accessions were chosen intentionally to represent different phenotypes and countries.

The most frequent haplotypic constitutions illustrate the biparental origin of this genetic pool (Motamayor et al. 2003) and corroborates our analytical rationale, highlighting ancestral haplotypes that represent chromosomal segments that were left intact through several generations of recombination (McKeigue 2005).

The extent of LD in crops is highly determined by the reproductive biology of the plant, outbreeders having generally less LD than inbreeders. However, there is considerable variation between various types of outbreeding materials. Traditional cultivars exhibit LD caused by early domestication bottlenecks, whereas modern cultivars display additionally the impact of recent breeding. The earliest example documented is that of sugarcane, whose modern varieties exhibit LD over distances of 10 cM, as the result of a bottleneck when breeding started from a few interspecific hybrids not many generations ago (Janoo et al. 1999). In maize, LD can decay within a few 100 bases in parts of the genome of landraces, but it can also span several 100 Kb in other genomic regions and populations (Remington et al. 2001). Strong LD has already been exploited in modern cultivars: Schnell et al. (2005) studying the parental origins of newly cultivated productive or unproductive hybrid cocoa trees from Hawaii, found microsatellite alleles associated with pod number. Kraakman et al. (2004) were able to localise 18–20 AFLP markers that accounted for 40–58% of the yield and yield stability in 146 modern two-row spring barley cultivars representing the current commercial germplasm in Europe; this study used data generated by commercial variety trials prior to variety release.

For seed and fruit weight, we found that both groups of cacao trees, plantation and collection, shared a higher number of associated regions with QTLs in other populations, than classical mapping progenies, showing the robustness of detected associations.

Associated regions identified in our target populations represent 56% of QTLs so far reported for the evaluated traits in separate mapping populations with a wider diversity of progenitors than our sampled individuals, with only Trinitario/Criollo origin. Besides, LD methodology was useful to detect additional marker-trait associations to those identified by QTL mapping studies.

The example of cacao takes advantage of earlier impacts of human activity, prior to conscious breeding schemes. We believe each crop, with its specific history of domestication and genetic improvement is amenable to such analysis, if the domestication process is clearly understood. The implications are important. Phenotypic information from the past characterisations can be used to derive valuable information about genomic structure and genetic control of useful traits without specifically developing mapping populations for QTL analyses, thus saving resources and increasing the scientific value of germplasm-related activities.

Association genetics or, like in this particular case, admixture mapping, can help curators and crop breeders to better understand the diversity held in germplasm collections, allowing them to conserve and use that diversity more effectively. The value added in this way to genebanks can help to encourage the investment of resources necessary to sustain such conservation efforts.

References

Aranzana JM, Kim S, Zhao K, Bakker E, Horton M, Jakob K, Lister C, Molitor J, Shindo C, Tang C, Toomajian C, Traw B, Zheng H, Bergelson J, Dean C, Marjoram P, Nordborg M. 2005. Genome-wide association mapping in Arabidopsis identifies previously known flowering time and pathogen resistance. PlosGenetics 1(5) www. plosgenetics.org

Benzecri JP (1973) L'analyse des données. Tome 2: L'analyse des correspondences. Dunod (eds), Paris, France

Belkhir K, Borsa P, Chikhi L, Raufaste N, Bonhomme F (2001) GENETIX 4.02, logiciel sous Windows TM pour la génétique des populations, Laboratoire Génome, Populations, Interactions: CNRS UMR. 5,000, Université de Montpellier II, Montpellier, France

Black WC, Krafsur ES (1985) A FORTRAN program for the calculation and analysis of two-locus linkage disequilibrium coefficients. Theor Appl Genet 70:491–491

Clément D, Risterucci AM, Motamayor JC, N'Goran J, Lanaud C (2003a) Mapping QTL for yield components, vigor and resistance to *Phytophthora palmivora* in *Theobroma cacao* L. Genome 46:103–111

Clément D, Risterucci AM, Motamayor JC, N'Goran J, Lanaud C (2003b) Mapping quantitative trait loci for bean traits and ovule number in *Theobroma cacao* L. Genome 46:204–212

Crouzillat D, Menard B, Mora A, Phillips W, Petiard V (2000) Quantitative trait loci analysis in *Theobroma cacao* L. using molecular markers. Euphytica 114:13–23

Engels JMM (1981) Technical bulletin 7, tropical agriculture and research training center, Turrialba, Costa Rica. 191 p

Darvasi A, Shifman S (2005) The beauty of admixture. Nat Genet 37(2):118–119

Excoffier L, Slatkin M (1995) Maximum likelihood estimation of molecular haplotypes frequencies in a diploid population. Mol Biol Evol 12(5):921–927

FAO (1998) The state of the world's plant genetic resources for food and agriculture, FAO (eds), Rome

Garnier-Gere P, Dillmann C (1992) LinkDos. J Hered 56:409–415

Guo SW, Thompson EA (1992) Performing the exact test of Hardy–Weinberg proportions for multiple alleles. Biometrics 48:361–372

Jannoo N, Grivet L, Dookun A, D'Hont A, Glaszmann JC (1999) Linkage disequilibrium among modern sugarcane cultivars. Theor Appl Genet 99:1053–1060

Kraakman ATW, Rients EN, Van den Berg PMMM, Stam P, Van Eeuwijk FA (2004) Linkage disequilibrium and mapping of yield and yield stability in modern spring barley cultivars. Genetics 168:436–444

Lanaud C, Boult E, Clapperton L, N'Goran NK, Cros E, Chapelin M, Risterucci, AM, Allaway D, Gilmour M, Cattaruzza A, Fouet O, Clément D, Petithugenin P (2003) Identification of QTLs related to fat content, seed size and sensorial traits in *Theobroma cacao* L. 14th International conference in cocoa, Accra, Ghana

McKeigue PM (2005) Prospects for admixture mapping of complex traits. Am J Hum Genet 76:1–7

Motamayor JC, Risterucci AM, Lopez PA, Lanaud C (2002) Cacao domestication I: the origin of the cacao cultivated by the Mayas. Heredity 89:380–386

Motamayor JC, Risterucci AM, Heath M, Lanaud C (2003) Cacao domestication II: progenitor germplasm of the Trinitario cacao cultivar. Heredity 91:322–330

Pritchard JK, Stephens M, Donnelly P (2000) Inference on population structure using multilocus genotype data. Genetics 155:945–959

Pugh T, Fouet O, Risterucci AM, Brottier P, Deletrez C, Courtois B, Clement D, Larmande P, N'Goran JAK, Lanaud C (2004) A new codominant markers based cocoa linkage map: Development and integration of 201 new microsatellites markers. Theor Appl Genet 108:1151–1161

Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR, Doebley J, Kresovich S, Goodman MM, Buckler ES (2001) Proc Nat Acad Sci USA 98:11479–11484

Rice WR (1989) Analysing tables of statistical tests. Evolution 43:223–225

Risterucci AM, Grivet L, N'Goran JAK, Pieretti I, Flament MH, Lanaud C (2000) A high-density linkage map of *Theobroma cacao* L. Theor Appl Genet 101:948–955

Schnell R, Olano CT, Brown JS, Meerow AW, Cervantes-Martinez C (2005) Retrospective determination of the parental population of superior cacao (*Theobroma cacao* L.) seedlings and association of microsatellite alleles with productivity. J Am Soc Hortic Sci 130(2):181–190

Thornsberry JM, Goodman MM, Doebley J, Kresowich S, Niel-
    sen D, Buckler ES (2001) *Dwarf8* polymorphisms associate
    with variation in flowering time. Nat Genet 28:286–289
Van Ooijen JW, Maliepaard C (1996) MapQTL (tm) Version 3.0:
    Software for the calculation of QTL positions on genetic
    maps, DLO-centre for plant breeding and reproduction re-
    search, Wageningen, The Netherlands
Wadsworth RM, Ford CS, Turnbull CJ, Hadley P (2003) Interna-
    tional cocoa germplasm Database v. 5.2., (Euronext. liffe/
    University of Reading, UK)

Weir BS, Cockerham CC (1979) Estimation of linkage disequilib-
    rium in randomly mating populations. Heredity 42:105–111
Weir BS (1996) Genetic data analysis II. Sinauer Associates Inc.
    Publishers, Sunderland, Massachusetts, 445 p
Zhu X, Luke A, Cooper RS, Quertermous T, Hanis C, Mosley T,
    Gu CC, Tang H, Rao DC, Risch N, Weder A (2005) Admix-
    ture mapping for hypertension loci with genome-scan mark-
    ers. Nat Genet 37:177–181